Neurocomputing 72 (2009) 2259–2269

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Using relevance feedback with short-term memory for content-based spine X-ray image retrieval

Xiaoqian Xu [a], Dah-Jye Lee [a,*], Sameer K. Antani [b], L. Rodney Long [b], James K. Archibald [a]

[a] Brigham Young University, Department of Electrical and Computer Engineering, Provo, UT, USA
[b] Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

ARTICLE INFO

Article history:
Received 2 April 2008
Received in revised form
29 November 2008
Accepted 20 December 2008
Communicated by T. Heskes
Available online 20 February 2009

Keywords:
Relevance feedback
Content-based image retrieval
Spine X-ray
Shape matching
Osteophytes

ABSTRACT

Managing large medical image databases has become a challenging task as more medical images are produced and stored in digital format. Computer-aided decision support for content-based image retrieval (CBIR) is an essential tool for medical image management. This paper presents a novel hybrid relevance feedback (RF) system for shape-based retrieval of spine X-ray images. A new shape similarity measure that considers both whole shape and partial shape matching is presented. The proposed RF architecture includes separate retrieval and feedback modes to solicit user's opinion for refining retrieval results. A unique short-term memory approach is implemented to avoid repeated request for user's feedback on the same, already approved, and retrieved relevant images. An automatic weight updating scheme is developed to present the images on which it is best for the user to provide feedback. Incorporating all these unique features, the proposed RF retrieval system is able to reduce the gap between high-level human visual perception and low-level computerized features. Experimental results show overall retrieval accuracy improvement of 22.0% and 17.5% after the second feedback iteration for retrieving spine X-ray images with similar osteophytes severity and type, respectively.

1. Introduction

Image retrieval [1–5] and content-based image retrieval (CBIR) [6–12] have been seen as a solution to managing large media collections. Research on medical image retrieval, however, is fairly recent [13–21]. With the increasing use of diagnostic medical imaging for training and research, efficient and effective image retrieval has become critical to medical image database management. Early CBIR research has focused on exploiting effective image features such as color, shape, and texture [6,7] or determining similarity [9–11]. Techniques extracting "high level" visual characteristics in the query image and computing similarity in terms of "low level" features result in a *semantic gap* or discrepancy between the two levels of representation. This semantic gap has been observed in the performance of most image features and similarity measurement methods used in CBIR systems. User feedback has been analyzed and employed to address this issue in image retrieval applications [22–28]. This issue, however, is more pronounced in medical image retrieval applications because medical images of the same anatomy but with different pathologies often exhibit very subtle differences that lead to different opinions even among experts.

Research efforts have sought to utilize CBIR methodologies to study a collection of 17,000 digitized spine X-ray images from the second National Health and Nutrition Examination Survey (NHANES II) which is maintained by the US National Library of Medicine (NLM) [29–33]. This collection is considered a key resource supporting research on the prevalence of osteoarthritis and musculoskeletal diseases. Our previous research efforts in CBIR techniques for spine X-ray images [29–31] have broadly focused on techniques for measuring shape similarity. Color and texture features often used in other CBIR systems are generally irrelevant for spine X-rays because they fail to capture the pathologies of interest. Although these shape matching techniques demonstrate promising results, they operate under the implicit assumption that a single set of fixed weights for a weighted similarity measurement is sufficient to express desired characteristics of various queries. Relevance feedback (RF) is considered a natural extension that could address this drawback as well as the common semantic gap issue mentioned previously.

The fundamental concept of RF is to establish interactions between the user and retrieval system and to refine retrieval based on feedback provided by the user. There are two types of image retrieval systems, *target search* and *category search*. A target search system searches for a specific target image in the database. A category search system, on the other hand, seeks a certain number of images that are similar to the query (target). Ideally, a category search system retrieves the images that are most similar

* Corresponding author.
E-mail address: djlee@ee.byu.edu (D.-J. Lee).

0925-2312/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2008.12.029

to the query's "class", e.g. pathology, type, or modality, which is our focus of this work.

There are many different RF algorithms published in the literature in recent years. Most of them were developed for specific applications. Probability estimation for calculating image similarity was proposed [21,25,34,35]. Feedback from the user was used to update the weights or re-estimate the probability. Bayesian rule was also used to estimate the probability of each image being the user's query [9,21,24]. The probability was conditioned on all feedback history from the user and updated globally during each feedback iteration [9]. The resulting system was quite sophisticated and the updating process was computationally expensive and proportional to the size of the image database. An alternative to this non-parametric Bayesian approach, expectation maximization (EM) has also been used to estimate statistical parameters such as the mean and the variance of the user's target distribution, given the assumption that the distribution is Gaussian [26]. The EM algorithm was applied to a maximum likelihood function chosen to make most images appear in the medium likelihood area.

Optimal adaptive learning is another approach employing RF that appears in the literature [10,21]. In [10], adaptive filters were used to imitate the human vision system. Least mean square and recursive least square algorithms were both proposed to approximate the optimal Wiener filter solution. The user's feedback was used as the ground truth to guide the algorithm to the optimal solution. Support vector machine (SVM) has also been used for active learning algorithms [4,5,21,36–39] to select the most informative images to query a user and learn the user's similarity preference for the next retrieval iteration [4]. Historical feedback data from the users have been used to learn effectively the correlation between low-level image features and high-level concepts using a so called soft labeled SVM [37].

Another important aspect of RF is image selection strategy that selects images for the user to provide feedback. Image selection strategy is actually crucial to the performance of a RF-based image retrieval system. Some approaches selected only images that were most similar to the user query for user feedback [14,22,24,25]. However, as the retrieval accuracy for a specific query increases from iteration to iteration, a large overlap between the selected image sets is inevitable and results in "over-learning". They also ignore useful information that can be obtained from the feedback regarding *negative images*, or irrelevant images retrieved by the system.

Generally speaking, RF is designed to bridge the semantic gap for enhancing performance. Users manually label positive and negative feedbacks to construct a classifier for later accurate retrieval [35,40]. Tao et al. [41] pointed out that the assumption of treating positive and negative feedbacks equally is not appropriate because two groups of training feedbacks have very different properties. They proposed an orthogonal complement analysis to solve this problem. Onoda et al. [39] proposed a unique one class SVM based method using only non-relevant information to retrieve relevant documents efficiently. A region segmentation based feedback process was proposed to estimate correspondence among regions in the feature space to separate relevant regions from irrelevant regions [42,43].

We proposed a novel linear weight-updating RF algorithm and applied it to spine X-ray image retrieval in [31]. Like most other RF approaches, new parameters (weights) intended to enhance the query expression are calculated after each feedback iteration, with refined retrieval results limited to those obtained by conventional CBIR using the new parameters. The work described herein contributes to spine X-ray image retrieval by proposing and evaluating a novel hybrid image retrieval approach based on CBIR and weighted shape matching that employs RF and feedback history.

This paper is organized as follows: Section 2 briefly introduces the shape similarity measure developed for vertebral image retrieval using RF. The improved linear weight-updating RF approach that employs short-term memory (STM) is introduced in Section 3. Section 4 details the hybrid approach of combining RF and CBIR for X-ray retrieval. Section 5 uses expert-established ground truth to evaluate the proposed algorithm, and the paper concludes in Section 6.

2. Shape similarity in spine X-ray images

Pathologies found in spine X-ray images that are of interest to the medical researchers are generally displayed along the vertebral boundary. These pathologies include anterior osteophytes (AO), intervertebral disc degeneration and resulting disc space narrowing, subluxation, and spondylolisthesis. Work presented in this paper focuses on AOs which are bony protrusions along the anterior, inferior and superior edges of the vertebra. These pathologies are expressed as protruding "corners" in the sagittal view. Fig. 1(a) shows a schematic, depicting landmark points of interest to medical researchers, including AOs, and Fig. 1(b) shows an example image from NHANES II collection that is cropped around a vertebra with superimposed boundary and landmark points.

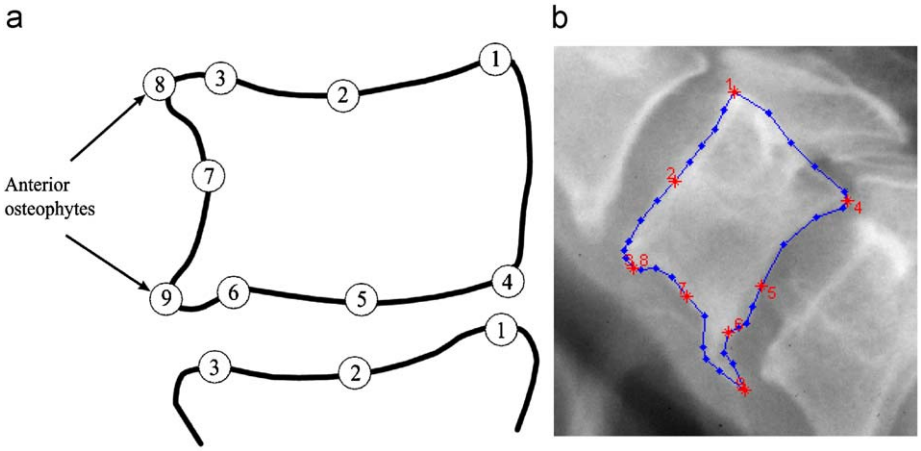

Fig. 1. Landmark points and a sample image: (a) radiologist-marked 9-point model, (b) an example image.

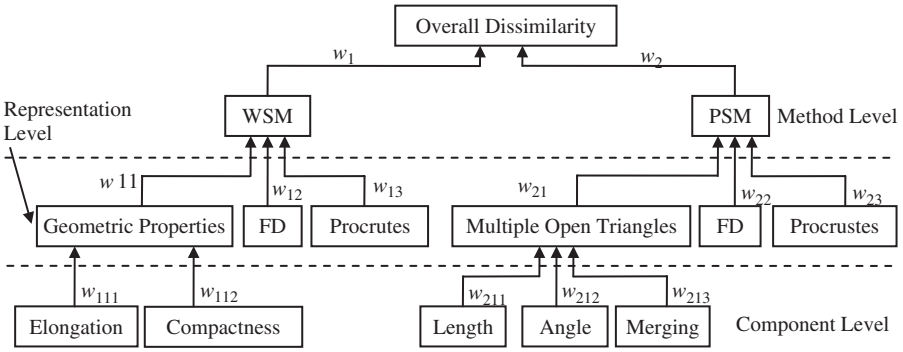

Fig. 2. Hierarchical retrieval model for spine shapes.

The proposed hierarchical retrieval model for our application is shown in Fig. 2. The overall dissimilarity between the query shape and a candidate shape in the database is determined along the two paths labeled WSM and PSM. WSM is a label for all similarity measures that match the whole shapes, while PSM identifies partial shape matching methods. As seen from bottom up, there are three hierarchical levels: the *component* level, the *representation* level, and the *method* level. One method can utilize multiple feature representations, and one representation can be computed from multiple components. For example, three representations are used by the WSM method: *Geometric properties*, *Fourier descriptor* (FD), and *Procustes* distance. The *geometric properties* representation consists of the two feature components *elongation* and *compactness*. A weight (w) is associated with each component, representation, and method, and the overall dissimilarity is calculated hierarchically as a weighted sum.

For clarity, the following expressions compute *dissimilarity*. (Similarity is often expressed as 1-*dissimilarity*). For example, for the PSM method, dissimilarity is calculated as

$$D_{PSM} = W_{21}D_{DP} + W_{22}D_{FD} + W_{23}D_{Pro} \quad (1)$$

where

$$D_{DP} = W_{211}D_{len} + W_{212}D_{ang} + W_{213}D_{mer} \quad (2)$$

D_{FD} is the L_2 distance between two fourier descriptor vectors representing two partial shapes and D_{Pro} is the Procrustes distance between two sets of shape data points. In Eq. (2), D_{len}, D_{ang}, and D_{mer} are the length dissimilarity, angle dissimilarity, and merging dissimilarity between two partial shapes, respectively. (A more detailed explanation of these dissimilarity measurements can be found in [31].) Similarly, dissimilarity for WSM methods is calculated as

$$D_{WSM} = W_{11}D_{GP} + W_{12}D_{FD} + W_{13}D_{Pro} \quad (3)$$

where

$$D_{GP} = W_{111}D_{elo} + W_{112}D_{com} \quad (4)$$

D_{elo} is the elongation dissimilarity between two whole shapes, while D_{com} is the compactness dissimilarity between two whole shapes. The overall dissimilarity is then calculated as the weighted sum:

$$D_{Overall} = W_1D_{WSM} + W_2D_{PSM} \quad (5)$$

Dissimilarities on each level are normalized to be in the range of (0, 1). Weights represent the relative importance of the corresponding component, representation, and method. These weights can be adjusted through RF to reflect user preference on different components, representations, and methods.

3. Relevance feedback

For shape-based retrieval systems like that described in the previous section, weight-updating approaches are preferred for relevance feedback.

3.1. Review and analysis

Like most statistical RF approaches, Rui's method [22] requires prior statistical information for all images in the database. During each iteration, the N objects that are most similar to the query are simply displayed to the user for feedback. According to his or her perception, the user assigns each of these N objects to one of the five categories: highly relevant, relevant, no opinion, non-relevant, or highly non-relevant. A numerical score (positive, negative, or zero) is then assigned to each image corresponding to its assigned category. Two different weight-updating approaches are taken for the feature and component levels. For the feature level, weights are updated by adding the newly assigned score to the original weights. For the component level, weights are updated as the reciprocal of the standard deviation of the component similarity value sequence from the relevant set specified by the user.

Given that the overall operation of our retrieval model is similar, it would seem reasonable to apply Rui's weight-updating scheme to our spine X-ray application. However, statistical information is difficult if not impossible to obtain for PSM, given the fact that there are unmanageably large number of possible partial queries that could be specified as query by the user. In addition, the dissimilarity measure used in PSM always depends on the specific query and thus has to be calculated on the fly.

Another issue is that it is common for one component that is able to differentiate the relevant set from the irrelevant set to have a larger deviation within the relevant set than a second component that cannot achieve the same level of differentiation. Therefore, weight-updating using the reciprocal of standard deviation is not ideal in some cases. Furthermore, weights for the feature level and the component level are updated independently, but this is both inefficient and ineffective. Suppose, for example, that a given feature does not perform well according to the user's feedback, and thus is assigned a lower weight during the weight-updating process. If the independently updated components of the feature happen to be assigned weights that make the feature a better indicator of the user perception, the logical conclusion is that the feature itself should have been assigned a higher weight.

3.2. Image selection for feedback

The proposed retrieval system has two modes: Mode R (Retrieval) and Mode F (Feedback). Mode R retrieves and displays

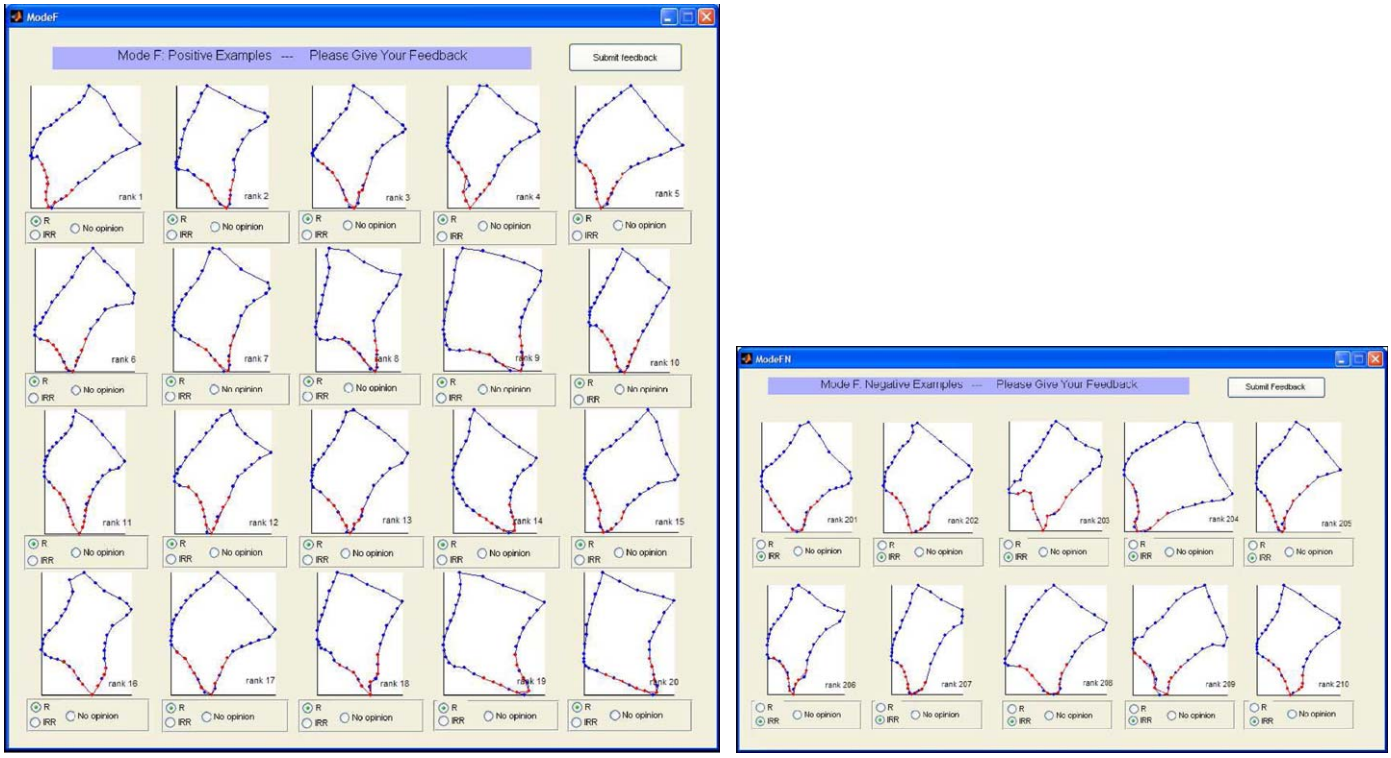

Fig. 3. Mode F: positive and negative examples.

the images most similar to the user's query, while Mode F displays the images and waits for user feedback. It is important to keep Mode R and Mode F independent since they employ different image selection schemes. The ultimate goal is to retrieve images that best match the user's query as well as preference. It is desired for the retrieval system to have a distinct "Mode R" to display the end results.

Fig. 3 shows 20 positive and 10 negative examples that await user feedback in Mode F. Fig. 4 illustrates the procedure for selecting examples for one specific query. The STM is refreshed when a new query is presented. Images in the database are ranked in Mode R in the order of increasing dissimilarity based on low-level image features. In the first feedback iteration, the top 20 matches from Mode R are all selected as positive examples for Mode F and stored in STM for recordkeeping. In subsequent iterations, only top matching images that have not been reviewed by the user in Mode F (and therefore not previously recorded in the STM) are stored in STM and output to Mode F for feedback. This approach ensures that top matching images in the database are selected as positive examples for feedback just once, thus avoiding repeated requests for user feedback on the same images.

Mode F also includes negative examples that are selected from images that have a high dissimilarity with the specific query. In our current system, 10 examples are chosen, beginning with the 200th ranked image in the last retrieval and continuing sequentially. As with the positive examples, the STM stores a record of negative examples shown to the user so that each is reviewed at most once. The inclusion of both positive and negative examples enhances the opportunity to make corrections in the event of severe misses on the positive matches while retaining sufficient feedback information on the positive matches.

3.3. *Weight-updating scheme*

The mechanism used to provide RF should be simple to allow the system to be used by practitioners not trained in CBIR concepts but capable of making visual judgments. In Mode F, the user is required only to express one of the three opinions for each image: the image is *relevant*, the user is *not sure*, or the image is *irrelevant*. The default opinion on positive examples is set to "relevant", while for negative examples it is set to "irrelevant".

We illustrate the operation of the new weight-updating scheme by describing it in detail at the component level (weights for the other two levels are updated in a similar fashion). First, assume that the dissimilarity values of a given component for the relevant and irrelevant sets are distributed as shown in Fig. 5. In an ideal case, the dissimilarity ranges for the relevant and irrelevant sets that do not overlap, which means that this component perfectly reflects the user's preference. Otherwise, for all the images in the relevant set, the range of dissimilarity values is obtained and denoted as; min_*R*, max_*R*. It is very likely that a dissimilarity value d_{IR} of one of the images from the irrelevant set obtained from user's feedback lies within the range (min_*R*, max_*R*). Thus, an ambiguous range of dissimilarity values occurs, as shown in Fig. 5. We define a difference related to this ambiguous range as:

$$Dif = d_{IR} - \max\_R \tag{6}$$

so that the sign of *Dif* indicates the occurrence of an ambiguous range between relevant image set and an image from the irrelevant image set in the following manner:

$$\begin{aligned} &Dif > 0 \quad \text{no ambiguity occurs} \\ &Dif < 0 \quad \text{an ambiguity occurs} \end{aligned} \tag{7}$$

note that only max_*R* is needed for calculating the overlapping range, since the dissimilarity values from irrelevant set usually would not go beyond min_*R*.

Suppose a representation consists of N components and the weights associated with these components are W_i, $i = 1,2,\ldots,N$. Every image marked as irrelevant by users is evaluated and for every occurrence of an ambiguous range in the component level,

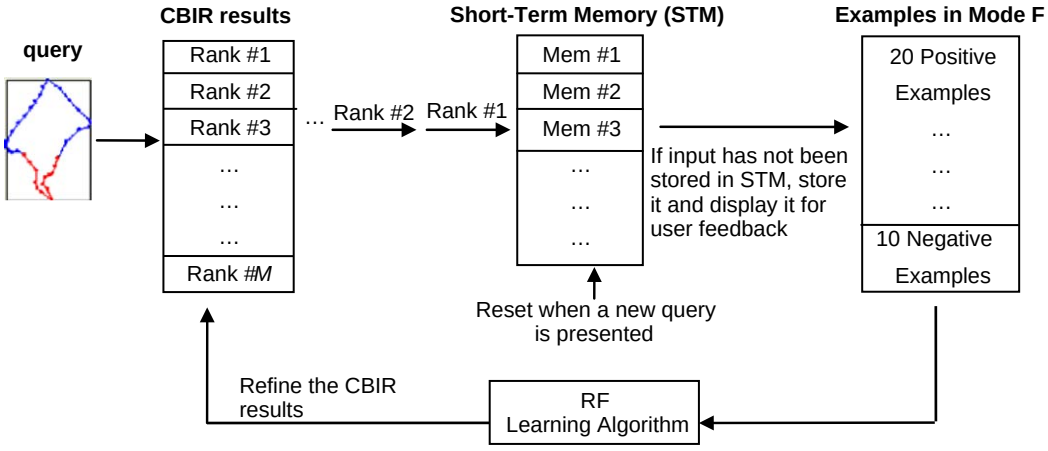

Fig. 4. Image selection scheme with short-term memory.

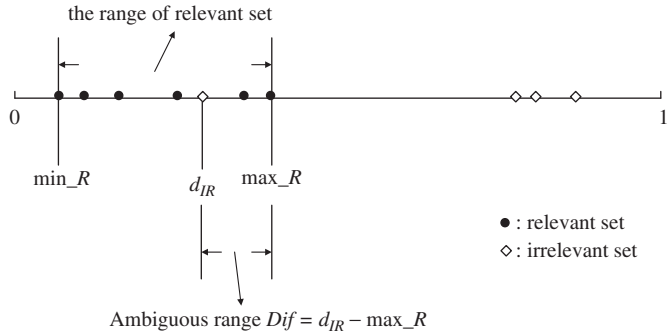

Fig. 5. The distribution of the dissimilarity values of a given component.

each weight is updated once according to

$$W_i = W_i + \max_{\substack{j \\ Dif_i - Dif_j > 0}} \left\{ \frac{Dif_i - Dif_j}{\max_{i,j}\{|Dif_i|, |Dif_j|\}} \frac{|W_i - W_j|}{\max\left\{W_i, \frac{1}{2}(W_i + W_j)\right\}} \right\} + \max_{\substack{j \\ Dif_i - Dif_j < 0}} \left\{ \frac{Dif_i - Dif_j}{\max_{i,j}\{|Dif_i|, |Dif_j|\}} \frac{|W_i - W_j|}{\max\left\{W_i, \frac{1}{2}(W_i + W_j)\right\}} \right\} \quad (8)$$

where Dif_i is calculated in the same way as in Eq. (6) but with the distribution of the dissimilarity values of a component in the current representation.

This approach updates the weights "dependently" by comparing the ambiguous ranges of different components, $Dif_i - Dif_j$, and the value of their corresponding weights, $|W_i - W_j|$. For instance, in order to update W_1, a sequence of values $Dif_1 - Dif_2$, $Dif_1 - Dif_3$,…, $Dif_1 - Dif_N$ are calculated. If $Dif_1 - Dif_j$ is positive, it indicates that the ith component creates a smaller ambiguous region (a bigger separation between the relevant and irrelevant sets) than the jth component. Therefore, the ith component deserves a higher weight than the jth component. In this case, our method calculates the positive value

$$\frac{Dif_1 - Dif_j}{\max\{|Dif_1|, |Dif_j|\}} \frac{|W_1 - W_j|}{\max\left\{W_1, \frac{1}{2}(W_1 + W_j)\right\}}$$

to be added to W_1. Instead of adding a random value (e.g., 1 or 2), the term $|W_i - W_j|$ serves as the base of weight-incremental unit so that the increment is at the level of existing weights. However, in order to provide a gradual increase in the weights and to avoid large fluctuations, a denominator $\max\{W_1, 1/2(W_1+W_j)\}$ is employed to provide a more stable incremental unit. The coefficient of the incremental unit is based on $Dif_1 - Dif_j$, and calculated as $(Dif_1 - Dif_j)/\max\{|Dif_1|, |Dif_j|\}$ in a similar manner as the incremental unit. The value based on the second term of Eq. (8) is calculated for each component which satisfies $Dif_1 - Dif_j > 0$. The maximum value is then chosen to be added to W_1. If $Dif_1 - Dif_j$ is negative, an analogical explanation holds for the third term in Eq. (8), except that a minimum negative value will be added to W_1.

Based on the above discussion, our weight-updating approach compares the ambiguous ranges of all components of a certain representation and updates the weights based on existing weight values and the sizes of the ambiguous regions. This dependent updating of weights is essential in the overall operation of our system; it ensures that the weights reflect the true importance of each component. The weights associated with the representation and method levels are updated using the same approach (as that of the component level discussed above) by observing every occurrence of an ambiguous range in the corresponding dissimilarity level.

A bottom-up approach is employed during the weight-updating procedure as shown in Fig. 6. Specifically, the weights of the components are updated first, and then the dissimilarity value of the corresponding representation is updated using the new weights for its components. Once the weights of all the components are updated according to the user feedback, the dissimilarity values of all the representations are updated as well. Therefore, instead of using the old dissimilarity values of the representations, the weights of the representations are updated by analyzing the new and updated dissimilarity values. This bottom-up approach addresses the disadvantages of Rui's method which updated the weights of all levels independently.

4. Hybrid approach based on CBIR and RF

4.1. Visual perception

As previously noted, in most RF approaches, the final retrieval results during any iteration come directly from the refined CBIR retrieval results. As noted in the literature (see [9,22,25,26]), this approach to incorporating RF provides a performance improvement

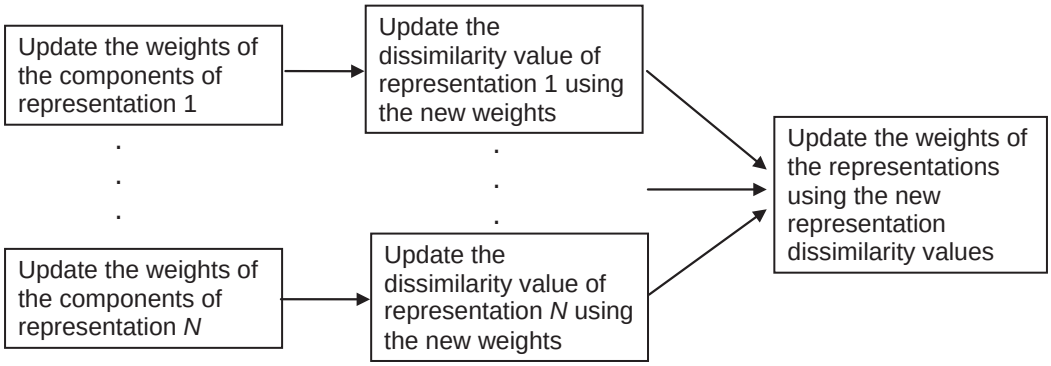

Fig. 6. Bottom-up updating procedure.

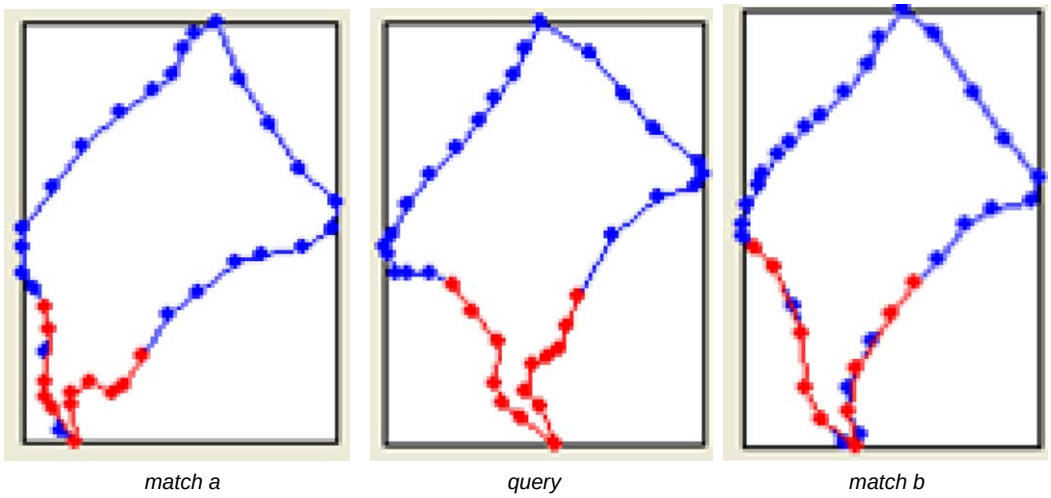

Fig. 7. Spine shapes.

over traditional CBIR techniques. However, this approach implicitly assumes that a single set of weights exists (for example, the set of all weights shown in Fig. 2 for calculating dissimilarity) for each query and for each user that can retrieve all the possible good matches to the query in the database. This assumption is not supported by the results of practical experiments.

The human vision system is very complicated, and the concept of "similarity" is subjective and imprecise. Research has explored what the concept of "similarity" means to humans [12,44–46]. Due to the subjective nature of the human vision system, different users can perceive the same image differently, with differences ranging from slight to significant. Moreover, the same user can perceive the same image from multiple aspects such as color or shape. Thus, a single set of weights (parameters) that favors either color or shape will not be able to retrieve images that match both criteria. Therefore, no matter how many feedback iterations the user attempts, typical RF algorithms will not refine the CBIR retrieval results to include both.

This argument is easily extended to medical images such as the spine X-ray images we consider, despite the fact that only shape is employed as a meaningful feature for spine X-ray image retrieval. Consider, for example, the three spine shapes in Fig. 7. If we consider the part of each shape highlighted in red as the region of interest, both matches a and b could be considered as good candidate matches to the query. Match a is similar to the query in its local angle characteristics but not the overall shape. In contrast, match b is a good match based on overall shape similarity but not based on local angle details. This underscores the multiple aspects employed by the human vision system. One set of optimal weights (parameters) is unlikely to reflect multiple aspects of visual perception.

4.2. Hybrid approach

Traditional RF approaches usually employ RF information only from the current iteration to refine the CBIR results, which become the retrieval results after each feedback iteration. In these schemes, RF from the current iteration (without STM) is fed directly into the weight-updating process and the final retrieval results are provided directly by the refined CBIR results. Thus, although they employ RF, traditional approaches are not able to accommodate the differences in perspective embodied in the human vision system as discussed in the previous section.

Feedback history has been employed in some previous designs to enhance the retrieval capabilities [37,47–48]. In [47], a high-dimensional feature vector was used to represent each lung image, and the images were initially retrieved based on an unweighted K nearest-neighbor method. A decision tree, which classifies each image to be either relevant or irrelevant, was constructed or trained with the knowledge of all RF history gained on the current query. The decision tree, in turn, returned all the images classified as relevant from which another K images closest to the query were retrieved for further feedback. Similarly, [37,48] employed feedback history for training purposes. In contrast, we propose to directly contribute feedback history to the refined retrieval results for each iteration.

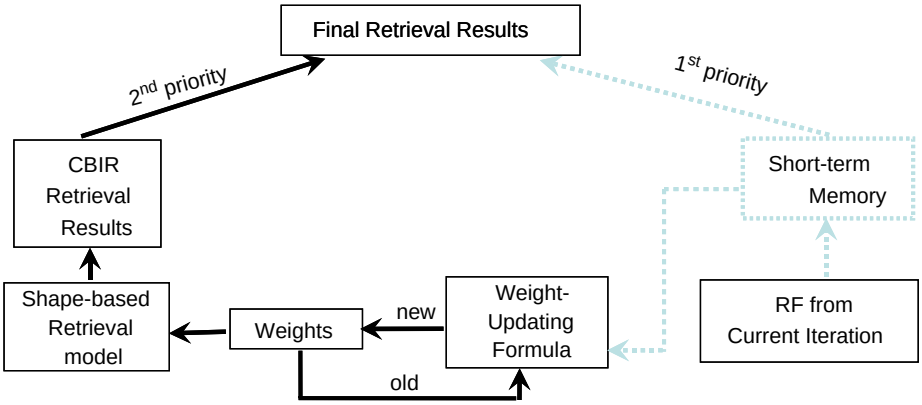

Fig. 8. Hybrid approach with short-term memory.

Table 1
AO severity grading criteria.

Severity	Slight	Moderate	Severe
Features	No narrowing or a $<15^\circ$ angle by the osteophyte from the expected normal anterior face of the vertebra or protrusion's length being $<1/5$ of vertebra width (traction) or height (claw)	Mild narrowing or an $(15^\circ, 45^\circ)$ angle by the osteophyte from the expected normal anterior face of the vertebra or protrusion's length being (1/5,1/3) of vertebra width (traction) or height (claw)	Sharp/severe narrowing or an $\geqslant 45^\circ$ angle by the osteophyte from the expected normal anterior face of the vertebra or protrusion's length being $>1/3$ of vertebra width (traction) or height (claw)

As the retrieval model in Fig. 8 shows, our proposed design includes STM that is reset at the beginning of each query. During each feedback iteration, the images selected for Mode F and the user's corresponding RF are recorded in the STM. Thus, instead of using the feedback information from only the current iteration, we use all available RF—including feedback from prior iterations stored in the STM—for the updating of weights. Since the STM for the image selection mechanism proposed in Section 3.2 stores all positive images selected for Mode F and records those images classified by the user as positive example matches, the same STM can also be used for the purpose of updating weights.

A new set of weights can be obtained during each iteration using the weight-updating formula proposed in Section 3.3 according to the RF information stored in the STM. The CBIR results are then refined using the new set of weights. Instead of using only the refined CBIR results as the final retrieval results, the final results are a combination of both the refined CBIR results and the positive RF stored in the STM (corresponding to the two paths entering the "final retrieval results" block in Fig. 8). Since the STM stores positive example matches that have been selected by the user, these images should be included as good matches in the final retrieval results. In fact, the top 20 matches in the final results could come from the STM if it contains 20 positive example matches for the current query. (This is somewhat unlikely; the user is likely to have been satisfied with the retrieval results before 20 images with positive feedback are stored in the STM.) This hybrid approach is more reasonable than typical RF approaches because it does not overlook the desirability, the efficiency, and arguably the necessity of including in the final retrieval results the best of all images retrieved by the system and approved by the user.

The intermediate images in the STM that received positive feedback from the user are likely from the results of different iterations. Since the weights are updated in each iteration, these images correspond to multiple, distinct set of weights. Therefore, by granting priority for these images to be included in the final retrieval results, our hybrid approach provides matches retrieved using multiple sets of weights rather just those retrieved using the final weights. The hybrid approach better reflects the multiple aspects of the human vision system.

5. Experimental results

5.1. Ground-truth establishment

The data set used to evaluate the effectiveness of the hybrid scheme consists of a total of 888 shapes generated from 207 spinal X-ray images (107 cervical and 100 lumbar films) selected from the NHANES II collection. Each vertebral shape boundary consists of 36 points consistently segmented with the first point at the posterior superior "corner" (Point 1 in the 9-point model). A set of 21 shapes were selected as queries for experiments.

To evaluate our approach, it was necessary to establish the ground truth and to interpret the shapes. Two classification schemes for AOs were chosen to establish the ground truth. One is the Macnab classification [49–51]. Two types of osteophytes are adapted from the Macnab classification: claw and traction. A claw spur rises from the vertebral rim and curves towards the adjacent disk. It is often triangular in shape and curved at the tips. A traction spur protrudes horizontally with moderate thickness, but does not curve at the tips or extend across the inter-vertebral disk space. The second classification is a severity grading system defined by medical experts and consistent with reasonable criteria for assigning severity levels to AO. Three severity levels are defined: slight, moderate, and severe. The criteria listed in Table 1 were developed based on [52].

By combining the two classification schemes, six categories of pathology can be established. Three examples of these categories are shown in Fig. 9. The ground truth of each selected X-ray image is based on the observations of a medical expert and recorded in a table. An example is shown in Fig. 10. For each shape, both the

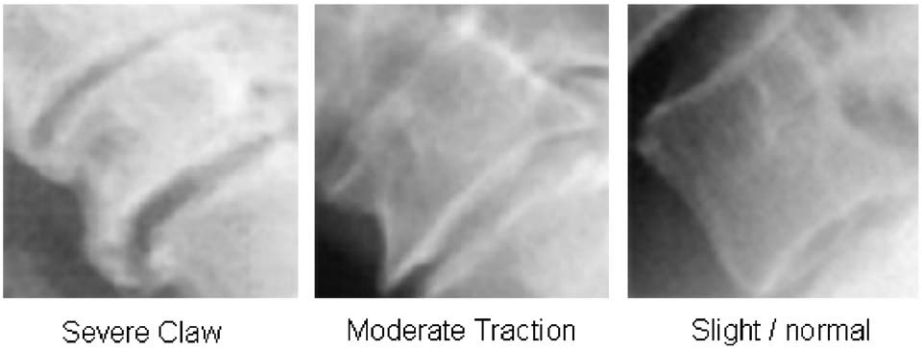

Fig. 9. Some examples of the Macnab classification and osteophyte severity grades.

ground_truth : Table

	filename	vertebra	location	claw	traction	grade
	C01235	C3	inferior	☑	☐	moderate
	C01235	C3	superior	☐	☐	slight
	C01235	C4	inferior	☑	☐	severe
	C01235	C4	superior	☐	☑	slight
	C01235	C5	inferior	☐	☐	slight

Record: 1 of 1901

Fig. 10. Ground truth table.

Table 2
Ground truth statistics.

	Inferior claw	Inferior traction	Superior claw	Superior traction
Severe	70 C/12 L	28 C/13 L	12 C/16 L	21 C/**3 L**
Moderate	92 C/**8 L**	53 C/26 L	**1 C**/12 L	20 C/31 L
Slight	79 C/24 L	24 C/91 L	3 C/2 L	57 C/58 L
Normal	89 C/295 L		318 C/337 L	

*C = Cervical, L = Lumbar.

anterior inferior and superior corners were classified and recorded separately. As with any medical diagnosis, the classification must be regarded as an opinion. While necessary for our evaluation, caution must be taken in considering this set as a gold standard. Ideally, a ground truth set should be developed through some form of consensus from multiple experts, taking observer variability into consideration. The development of a large ground truth set is an ongoing project at NLM.

Once the ground truth was established, it became clear that the six categories of pathologies do not occur with the same frequency in the database. Most of the shapes in the database belong to the "slight" severity categories, and there are very few samples for several other categories. A complete summary of the category distribution for the ground truth data is given in Table 2. "Normal" samples are those classified to be of slight severity without any of the Macnab types. Because of the subtle differences within the "slight" severity level, normal shapes are considered to be in the same category as any "slight" shape, although a slight claw shape is considered to be in a different category than a slight traction shape. In Table 2, the categories which have fewer than 10 samples in the database are in boldface, and those having between 10 and 19 samples are in italics. Regardless of severity, superior claw appears rarely in the database.

5.2. Experimental results and performance evaluation

Based on Table 1, 24 queries were selected for evaluation. Each query was a unique combination of severity level (slight, moderate, or severe), Macnab type (claw or traction), osteophyte location (superior or inferior), and image types (cervical or lumbar). However, queries belonging to the three boldfaced categories in Table 2 were excluded from the experiments because of their low representation in the database. For all 21 remaining queries, we performed two independent sets of evaluations as shown in Table 3, based on either severity or type. For the severity column, a shape was considered to be a good match if it had the same severity level as the query according to the ground truth. Therefore, during the RF process, such shapes were marked as "relevant" and all others were marked "irrelevant". In this case, the RF from the user was independent of Macnab type. The corresponding approach was taken to obtain the results in the Type column: retrieved images matching the type of the query were generally classified as "relevant", and others were deemed to be "irrelevant". However, normal shapes were marked as "relevant" to any slight shape, even though slight claw shape was still considered "irrelevant" to slight traction shape in the query.

For both sets of testing, i.e. severity and type, up to two iterations of RF were conducted. For each query, the top 20 matches were retrieved for study. The *retrieval accuracy percentage* was calculated for each feedback iteration. In the Severity column, the accuracy is defined as the percentage of the shapes among the top 20 matches with the same severity level as the query; in the Type column, the accuracy is defined as the

Table 3
Accuracy results: hybrid approach.

	Severity (RF insensitive to type, position, and location)				Type (RF insensitive to severity, position, and location)		
	Severe (%)	Moderate (%)	Slight (%)	21 queries (%)	Claw (%)	Traction (%)	21 queries (%)
Without feedback	47.14	48.33	85.00	60.75	74.44	79.55	77.25
After 1st feedback iteration	55.71	62.50	97.14	72.25	82.78	90.45	87.00
After 2nd feedback iteration	68.57	79.17	100	82.75	88.89	99.55	94.75
Overall improvement	21.43	30.84	15.00	22.00	14.45	20.00	17.50

Table 4
Accuracy results: hybrid approach without "Negative Examples" in RF.

	Severity (RF insensitive to type, position, and location)				Type (RF insensitive to severity, position, and location)		
	Severe (%)	Moderate (%)	Slight (%)	21 queries (%)	Claw (%)	Traction (%)	21 queries (%)
Without feedback	47.14	48.33	85.00	60.75	74.44	79.55	77.25
After 1st feedback iteration	57.43	62.50	97.86	73.75	82.33	88.59	85.50
After 2nd feedback iteration	65.71	75.83	99.29	80.50	88.39	98.59	93.50
Overall improvement	18.57	27.50	14.29	19.75	13.95	19.04	16.25

Table 5
Accuracy results: original weight-updating approach.

	Severity (RF insensitive to type, position, and location)				Type (RF insensitive to severity, position, and location)		
	Severe (%)	Moderate (%)	Slight (%)	21 queries (%)	Claw (%)	Traction (%)	21 queries (%)
Without feedback	47.14	48.33	85.00	60.75	74.44	79.55	77.25
After 1st feedback iteration	54.83	58.50	92.43	64.50	79.03	83.27	81.75
After 2nd feedback iteration	59.47	66.75	97.29	73.68	84.49	91.33	88.23
Overall improvement	12.33	18.42	12.29	12.93	10.05	11.78	10.98

percentage of the shapes in the top 20 matches with the same Macnab type as the querys. In both the severity and type columns, the results are presented for each category as well as for all 21 queries. For instance, in the severity column, the average accuracy percentages are shown for all of the severe, moderate, slight, and 21 queries, respectively. Similarly, in the Type column, the accuracy results are calculated for all the claw, traction, and 21 queries, respectively.

For comparison, we conducted experiments independently using the following three approaches: the proposed hybrid approach, the hybrid approach without "negative examples" in the RF process, and our original approach based solely on our weight-updating scheme. Experimental results are shown in Tables 3–5, respectively. In both sets of experiments (severity and type), the proposed hybrid approach showed significant improvements in just two feedback iterations. The overall improvement for the severity test was 22.00% with 82.75% accuracy percentage after the second feedback iteration. The overall improvement for the type test was 17.50% with a high 94.75% accuracy percentage after the second feedback iteration.

When negative examples are excluded from the RF process, Table 4 shows a slightly lower improvement after the second feedback iteration compared with the results in Table 3. However, in severity tests, the exclusion of "negative examples" seemed to produce higher retrieval accuracy after the first iteration. Therefore, more experiments are needed to determine conclusively the role of negative examples. The new hybrid approach clearly outperformed the original weight-updating approach, as a comparison of Tables 2 and 5 makes clear. The overall improvement percentage of the hybrid approach was almost twice of that of the original weight-updating method.

It is worth noticing that the accuracy percentage was inevitably reduced due to the fact that some categories (italicized in Table 2) have fewer than 20 samples in the database. To some extent, this explains why the overall accuracy percentage for severe queries is lower than that of slight qeuries. In some cases, RF showed no improvement in the second or even the first feedback iteration because of the lack of sufficient samples. Thus, we expect that our results would improve if we were able to test our approach on a more comprehensive database. Further experiments will be performed once a large ground truth set is established at NLM.

The precision measured here is for the retrieval system as a whole, which is the combination of the matching algorithm and RF process, rather than the matching algorithm alone. Because of the nature of the proposed hybrid approach, i.e. the number of

retrieved relevant images increases from iteration to iteration with user feedback, the "precision" and "recall" will in fact increase in a similar manner. Although there exists a more complete method to evaluate the performance of a retrieval system [53], using only "precision" measurement for this evaluation is sufficient.

6. Conclusion

In this paper, we present a new relevance feedback (RF) approach for building an image retrieval system for the second National Health and Nutrition Examination Survey (NHANES II) spine X-ray database that is maintained by the US National Library of Medicine (NLM). This novel hybrid approach directly utilizes feedback history and the modified CBIR results through RF. A new RF technique using short-term memory (STM) to store feedback history is developed. Through the use of an STM, we also propose an image selection scheme that significantly improves the efficiency and effectiveness of soliciting user's RF. A new weight-updating method which analyzes feedback information from the user to refine CBIR results is also introduced. We discuss human visual perception to demonstrate the advantages of this hybrid approach. For the two sets of experiments (severity and type), the overall improvement was 19.75% with an 88.75% average accuracy percentage. We achieved 100% retrieval accuracy for six queries which have sufficient representations in the database. Our approach offers significant potential for a shape-based medical image retrieval system requiring user interaction.

Building a more comprehensive database is a primary goal of future work. More experiments can be conducted with a better database and reliable ground truth to evaluate our new approach. In addition, combining CBIR with semantic information (i.e., classification labels such as severity and type) will likely improve retrieval accuracy even further. Our ultimate goal, of course, is to integrate this new approach into a web-based spine X-ray image retrieval system maintained by NLM.

Acknowledgments

This work research was supported in part by the National Library of Medicine (NLM) under Grant contracts #467-MZ301975 and #467-MZ401638 and intramural research funds of the Lister Hill National Center for Biomedical Communications, the National Library of Medicine (NLM), and the National Institutes of Health (NIH).

References

[1] K. Kuroda, M. Hagiwara, An image retrieval system by impression words and specific object names—IRIS, Neurocomputing 43 (4) (2002) 259–276.
[2] H. Frigui, Interactive image retrieval using fuzzy sets, Pattern Recognition Letters 22 (9) (2001) 1021–1031.
[3] S.-M. Hsieh, C.-C. Hsu, Graph-based representation for similarity retrieval of symbolic images, Data & Knowledge Engineering 65 (3) (2008) 401–418.
[4] S. Tang, E. Chang, Support vector machine active learning for image retrieval, in: ACM International Conference on Multimedia, 2001, pp. 107–118.
[5] R. Liu, Y. Wang, T. Baba, D. Masumoto, S. Nagata, SVM-based active feedback in image retrieval using clustering and unlabeled data, Pattern Recognition 41 (8) (2008) 2645–2655.
[6] S. Deb, Y. Zhang, An overview of content-based image retrieval techniques, in: Proceedings of the 18th International Conference on Advanced Information Networking and Applications, vol. 1, 2004, pp. 59–64.
[7] A. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 1349–1380.
[8] P. Muneesawang, L. Guan, A neural network approach for learning image similarity in adaptive CBIR, IEEE Fourth Workshop on Multi-media Signal Processing (2001) 257–262.
[9] I. Cox, M. Miller, T. Minka, T. Papathomas, P. Yianilos, The Bayesian image retrieval system, Pichunter: theory, implementation, and psychophysical experiments, IEEE Transactions on Image Processing 9 (2000) 20–37.
[10] T. Wang, Y. Rui, S.M. Hu, Optimal adaptive learning for image retrieval, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2001, pp. 1140–1147.
[11] I. El-Naqa, Y. Yang, N. Galatsanos, R. Nishikawa, M. Wernick, A similarity learning approach to content-based image retrieval: application to digital mammography, IEEE Transactions on Medical Imaging 23 (2004) 1233–1244.
[12] C.Y. Chiu, H.C. Lin, S.N. Yang, A fuzzy logic CBIR system, in: The 12th IEEE International Conference on Fuzzy Systems, vol. 2, 2003, pp. 1171–1176.
[13] H.D. Tagare, C.C. Jaffe, J. Duncan, Medical image databases: a content-based retrieval approach, Journal of the American Medical Informatics Association 4 (1997) 184–198.
[14] C.J.S. Ghebreab, A.W. Smeulders, Concept-based retrieval of biomedical images, in: Proceedings of SPIE Medical Imaging: PACS and Integrated Medical Information Systems, vol. 5033, 2003, pp. 97–108.
[15] M.M. Rahman, P. Bhattacharya, B.C. Desai, A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback, IEEE Transactions on Information Technology in Biomedicine 11 (1) (2007) 58–69.
[16] J. Kim, W. Cai, D. Feng, H. Wu, A new way for multidimensional medical data management: volume of interest (VOI)-based retrieval of medical images with visual and functional features, IEEE Transactions on Information Technology in Biomedicine 10 (3) (2006) 598–607.
[17] A. Mojsilovic, J. Gomes, Semantic based categorization, browsing and retrieval in medical image databases, International Conference on Image Processing 3 (2002) 145–148.
[18] H. Greenspan, A.T. Pinhas, Medical image categorization and retrieval for PACS using the GMM-KL framework, IEEE Transactions on Information Technology in Biomedicine 11 (2) (2007) 190–202.
[19] T.M. Lehmann, B.B. Wein, D. Keysers, M. Kohnen, H. Schubert, A monohierarchical multiaxial classification code for medical images in content-based retrieval, IEEE International Symposium on Biomedical Imaging (2002) 313–316.
[20] W. Liu, Q.Y. Tong, Medical image retrieval using salient point detector, in: IEEE-EMBS 27th Annual International Conference of the Engineering in Medicine and Biology Society, 2005, pp. 6352–6355.
[21] M. Rahman, B.C. Desai, P. Bhattacharya, Medical image retrieval with probabilistic multi-class support vector machine classifiers and adaptive similarity fusion, Computerized Medical Imaging and Graphics 32 (2) (2008) 95–108.
[22] Y. Rui. T. Huang, M. Ortega, S. Mehrotra, Relevance feedback: a power tool for interactive content-based image retrieval, IEEE Transactions on Circuits and Systems for Video Technology 8 (1998) 644–655.
[23] G.D. Guo, A. Jain, W.Y. Ma, H.J. Zhang, Learning similarity measure for natural image retrieval with relevance feedback, IEEE Transactions on Neural Networks 13 (2002) 811–820.
[24] I.J. Cox, M.L. Miller, S.M. Omohundro, P.N. Yianilos, Pichunter: Bayesian relevance feedback for image retrieval, International Conference on Pattern Recognition 3 (1996) 362–369.
[25] I. King, Z. Jin, Integrated probability function and its application to content-based image retrieval by relevance feedback, Pattern Recognition 36 (2003) 2177–2186.
[26] K. Sia, I. King, Relevance feedback based on parameter estimation of target distribution, International Joint Conference on Neural Networks 2 (2002) 1974–1979.
[27] S.K. Saha, A.K. Das, B. Chanda, Image retrieval based on indexing and relevance feedback, Pattern Recognition Letters 28 (3) (2007) 357–366.
[28] T. Zhao, L.H. Tang, H.H.S. Ip, F. Qi, On relevance feedback and similarity measure for image retrieval with synergetic neural nets, Neurocomputing 51 (2003) 105–124.
[29] L.R. Long, S. Antani, G. Thoma, Image informatics at a national research center, Computerized Medical Imaging and Graphics 29 (2) (2005) 171–193.
[30] S. Antani, D.J. Lee, L.R. Long, G. Thoma, Evaluation of shape similarity measurement methods for spine X-ray images, Journal of Visual Communication and Image Representation 15 (2004) 285–302.
[31] X.Q. Xu, D.J. Lee, S.K. Antani, L.R. Long, A spine X-ray image retrieval system using partial shape matching, IEEE Transactions on Information Technology in Biomedicine 12 (1) (2008) 100–108.
[32] T.M. Deserno, S. Antani, R. Long, Ontology of gaps in content-based image retrieval, Journal of Digital Imaging, doi: 10.1007/s10278-007-9092, 2007.
[33] J. Yao, Z. Zhang, S. Antani, R. Long, G. Thoma, Automatic medical image annotation and retrieval, Neurocomputing 71 (10–12) (2008) 2012–2022.
[34] M.L. Kherfi, D. Ziou, Relevance feedback for CBIR: a new approach based on probabilistic feature weighting with positive and negative examples, IEEE Transactions on Image Processing 15 (4) (2006) 1017–1030.
[35] T. León, P. Zuccarello, G. Ayala, E. de Ves, J. Domingo, Applying logistic regression to relevance feedback in image retrieval systems, Pattern Recognition 40 (10) (2007) 2621–2632.
[36] I. El-Naqa, Y. Yang, N. Galatsanos, M. Wernick, Relevance feedback based on incremental learning for mammogram retrieval, IEEE International Conference on Image Processing 1 (2003) 729–732.
[37] S.C.H. Hoi, M.R. Lyu, R. Jin, A unified log-based relevance feedback scheme for image retrieval, IEEE Transactions on Knowledge and Data Engineering 18 (4) (2006) 509–524.

[38] M. Ferecatu, N. Boujemaa, Interactive remote-sensing image retrieval using active relevance feedback, IEEE Transactions on Geoscience and Remote Sensing 45 (4) (2007) 818–826.
[39] T. Onoda, H. Murata, S. Yamada, Non-relevance feedback document retrieval based on one class SVM and SVDD, in: IEEE International Joint Conference on Neural Networks, 2006, pp. 1212–1219.
[40] M.L. Kherfi, D. Ziou, A. Bernardi, Combining positive and negative examples in relevance feedback for content-based image retrieval, Journal of Visual Communication and Image Representation 14 (4) (2003) 428–457.
[41] D. Tao, X. Tang, X. Li, Which components are important for interactive image searching?, IEEE Transactions on Circuits and Systems for Video Technology 18 (1) (2008) 3–11.
[42] C.Y. Li, C.T. Hsu, Image retrieval with relevance feedback based on graph-theoretic region correspondence estimation, IEEE Transactions on Multimedia 10 (3) (2008) 447–456.
[43] J.W. Kwak, N.I. Cho, Relevance feedback in content-based image retrieval system by selective region growing in the feature space, Signal Processing: Image Communication 18 (9) (2003) 787–799.
[44] C.Y. Chiu, H.C. Lin, S.N. Yang, Learning human perceptual concepts in a fuzzy CBIR system, in: The Fifth International Conference on Computational Intelligence and Multimedia Applications, vol. 2, 2003, pp. 330–335.
[45] A. Majumdar, I. Bhattacharya, A. Saha, An object-oriented fuzzy data model for similarity detection in image databases, IEEE Transactions on Knowledge and Data Engineering 14 (2002) 1186–1189.
[46] L.B. Chen, Y.N. Wang, B.G. Hu, Kernel-based similarity learning, in: Proceedings of International Conference on Machine Learning and Cybernetics, vol. 4, 2002, pp. 2152–2156.
[47] S. MacArthur, C.E. Brodley, A.C. Kak, L.S. Broderick, Interactive content-based image retrieval using relevance feedback, Computer Vision and Image Understanding 88 (2002) 55–75.
[48] H. Wang, B.C. Ooi, A.K.H. Tung, iSearch: mining retrieval history for content-based image retrieval, The Eighth IEEE International Conference on Database Systems for Advanced Applications, 2003, pp. 275–282.
[49] M. Heggeness, B. Doherty, Morphologic study of lumbar vertebral osteophytes, Southern Medical Journal 91 (1998) 187–189.
[50] D. Pate, J. Goobar, D. Resnick, P. Haghighi, D. Sartoris, M. Pathria, Traction osteophytes of the lumbar spine: radiographic-pathologic correlation, Radiology 166 (1988) 843–846.
[51] I. Macnab, The traction spur: an indicator of segmental instability, Journal of Bone and Joint Surgery 53 (1971) 663–670.
[52] J. Shou, S. Antani, L.R. Long, G. Thoma, Evaluating partial shape queries for pathology-based retrieval of vertebra, in: Proceedings of the Eighth World Multiconference on Systemics, Cybernetics and Informatics, vol. 12, 2004, pp. 155–160.
[53] D.P. Huijsmans, N. Sebe, How to complete performance graphs in content-based image retrieval: add generality and normalize scope, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2) (2005) 245–251.

Xiaoqian Xu received her B.S. degree in electrical engineering from Najing University, China in 2002. She then received her Ph.D. degree in electrical engineering from Brigham Young University, Provo, UT, in December 2006.

During her Ph.D. studies, she worked as a summer intern at the National Institutes of Health during the summers of 2003 and 2005. Her Ph.D. research topics included Content-Based Image Retrieval, relevance feedback, and indexing with applications to medical image retrieval. She is currently working as an applications engineer at KLA-Tencor Corporation in San Jose, CA.

Dah-Jye Lee received his B.S.E.E. from National Taiwan University of Science and Technology in 1984, M.S. and Ph.D. degrees in electrical engineering from Texas Tech University in 1987 and 1990, respectively. He also received his MBA degree from Shenandoah University, Winchester, Virginia, in 1999.

He is currently an Associate Professor in the Department of Electrical and Computer Engineering at Brigham Young University. He worked in the machine vision industry for eleven years prior to joining BYU in 2001. His research work focuses on Medical informatics and imaging, shape-based pattern recognition, hardware implementation of real-time 3-D vision algorithms and machine vision applications.

Dr. Lee is a senior member of IEEE and a member of SPIE. He has actively served the research community as a paper and proposal reviewer and conference organizer.

Sameer Antani received his B.E. (Computer) from University of Pune in 1994, M.E. and Ph.D. degrees in Computer Science and Engineering from the Pennsylvania State University in 1998 and 2001, respectively.

He is currently a Staff Scientist with the Lister Hill National Center for Biomedical Communications an intramural R&D division of the National Library of Medicine which is an institute within the US National Institutes of Health. His research interests are in data management of and retrieval from, large biomedical multimedia archives. His research includes content-based indexing, and retrieval of biomedical images, combining image and text retrieval, and next-generation documents that are enriched with interconnections to data sets and multimedia content.

Dr. Antani is a member of IEEE and IEEE Computer Society. He serves on the steering committee for IEEE Symposium for Computer Based Medical Systems (CBMS). He is a reviewer for several journals including various IEEE transactions.

L. Rodney Long received his B.A. and M.A. degrees in mathematics from the University of Texas in 1971 and 1976, respectively, and M.A. degree in applied mathematics from the University of Maryland in 1987.

Since 1990 he has been an Electronics Engineer in the Communications Engineering Branch of the National Library of Medicine. He previously worked for 14 years in industry as a programmer and engineer. His research work is concerned with Content-Based Image Retrieval, image processing, and image databases for biomedical applications.

He has been a member of IEEE since 1986 and has served as co-chair of the IEEE International Symposium of Computer-Based Medical Systems.

James K. Archibald received the B.S. degree in mathematics from Brigham Young University in 1981 and the M.S. and Ph.D. degrees in computer science from the University of Washington in 1983 and 1987, respectively. He has been with the Electrical and Computer Engineering Department at Brigham Young University since 1987. His research interests include robotics, multi-agent systems, and machine vision. Dr. Archibald is a member of the ACM and Phi Kappa Phi.